A Spectral Approach for Probabilistic Grammatical Inference on Trees

Raphaël Bailly, Amaury Habrard, Francois Denis

A Spectral Approach for Probabilistic Grammatical Inference on Trees[*]

Raphaël Bailly, Amaury Habrard, and François Denis

Laboratoire d'Informatique Fondamentale de Marseille,
UMR CNRS 6166, Aix-Marseille Université
CMI, 39 rue F. Joliot Curie, 13453 Marseille cedex 13, France
{raphael.bailly,amaury.habrard,francois.denis}@lif.univ-mrs.fr

Abstract. We focus on the estimation of a probability distribution over a set of trees. We consider here the class of distributions computed by weighted automata - a strict generalization of probabilistic tree automata. This class of distributions (called rational distributions, or rational stochastic tree languages - RSTL) has an algebraic characterization: All the residuals (conditional) of such distributions lie in a finite-dimensional vector subspace. We propose a methodology based on Principal Components Analysis to identify this vector subspace. We provide an algorithm that computes an estimate of the target residuals vector subspace and builds a model which computes an estimate of the target distribution.

1 Introduction

In this article, we focus on the problem of learning probability distributions over trees. This problem is motivated by the high need in XML applications or natural language processing to represent large tree sets by probabilistic models. From a machine learning standpoint, this problem can be formulated as follows. Given a sample of trees independently drawn according to an unknown distribution p, a classical problem is to infer an estimate of p in some class of probabilistic models [1]. This is a classical problem in grammatical inference and the objective here is to find a good estimate of the model's parameters. A usual class of models is the class of probabilistic tree automata (PTA) where the parameters lie in $[0, 1]$.

Recent approaches propose using a larger class of representation: the class of rational distributions (also called rational stochastic tree languages, or RSTL) that can be computed by weighted tree automata - with parameters in $\mathbb{R}$, hence with weights that can be negative and without any per state normalisation condition. This class has two interesting properties: It has a high level of expressiveness since it strictly includes the class of PTA and it admits a canonical form with a minimal number of parameters (see [2] for an illustration in the string case). It has notably the characterization that the residuals of a rational distribution (a special kind of conditional distributions) lie in a

finite-dimensional subspace. This set of residuals spans a vector subspace W of the vector space of real values functions over trees. W is finite dimensional and its dimension corresponds to the minimal number of states needed by a weighted tree automaton to compute p. Thus, a goal of an inference algorithm might be to identify this subspace W. This was illustrated by the algorithm DEES [3,4] which builds iteratively a weighted automaton computing an estimate of p. However, the iterative approach presented before suffers from the drawback to rely on statistical tests that are done on fewer and fewer examples when the structure grows.

In order to overcome this drawback, in this paper we investigate the possibility of using Principal Component Analysis (PCA) to identify the target vector subspace spanned by the residuals of a rational distribution, and then to build a representation from this subspace. PCA has already been used in grammatical inference for learning rational string distributions in [5], and in another framework in [6]. Another spectral approach was proposed in [7,8] for learning a class of Hidden Markov Models (HMM) over sequences. In this paper, we show that considering the class of rational distributions offers a natural framework for applying PCA to identify the target residuals subspace. Moreover, we obtain a high gain of expressiveness since we are able to infer classes of distributions that can not be computed by PTA. This gain in expressiveness has unfortunately two main drawbacks: the class of rational distributions is not recursively enumerable and it is not decidable if a rational series defines a distribution [9]. In spite of these strong constraints, we give some asymptotic error bounds and provide pointwise convergence result.

The paper is organized as follows. Section 2 gives the preliminaries on trees and rational tree series. Section 3 is devoted to our algorithm, while the convergence properties are presented in Section 4. Some experiments are provided in the last section.

2   Preliminaries

In this section, we introduce the objects that will be used all along in the paper. We mainly follow notations and definitions from [10] about trees. Formal power tree series have been introduced in [11] where the main results appear. Some notations about norms and matrices terminate this section.

2.1   Trees and Contexts

Unranked Trees Let F be an unranked alphabet. The set of *unranked trees* over F is the smallest set T_F satisfying $F \subseteq T_F$, and for any $f \in F$, and $t_1, \ldots, t_m \in T_F$, $f(t_1, \ldots, t_m) \in T_F$.

Ranked Trees Let $F = F_0 \cup \cdots \cup F_n$ be a ranked alphabet where the elements in F_0 are also called constant symbols. The set of *trees* over F is the smallest set T_F satisfying $F_0 \subseteq T_F$, and for any $f \in F_k$, and any $t_1, \ldots, t_k \in T_F$, $f(t_1, \ldots, t_k) \in T_F$.

Any tree defined over an unranked alphabet F can be represented over a ranked alphabet $F^@ = F_2^@ \cup F_0^@$ with only one binary symbol @, i.e. $F_2^@ = \{@(\cdot, \cdot)\}$ where $@ \notin F$ and constants that comprise all symbols in F: $F_0^@ = F$. Figure 1(d) shows such a representation (called curryfication) of the tree of Figure 1(c). Curryfication can be formally defined by induction:

- $curry(f(t_1, \ldots, t_n)) = @(curry(f(t_1, \ldots, t_{n-1})), curry(t_n))$
- $curry(f(t)) = @(f, curry(t))$
- $curry(a) = a$ for $a \in F$

This particular class of ranked alphabet is in bijection with the set of *unranked trees* [10], i.e. labeled trees in which any node may have an unbounded number of children. Weighted automata on unranked trees are defined in [12], where it is proved that weighted unranked tree automata on F are equivalent to weighted tree automata on $F^@$. As ranked trees are a particular case of unranked trees and weighted ranked tree automata can be seen as a particular case of weighted unranked tree automata, the results still hold for any ranked alphabet.

Hence, without loss of generality, and in all the rest of the paper, we will only consider a ranked alphabet equipped with constant symbols and with only one binary symbol in the following of the paper. For convenience, we will use f for denoting the binary symbol instead of @.

Contexts Contexts are element c of $C_F \subset T_{F \cup \{\$\}}$ where \$ is a variable that appears exactly once as a leaf in c (\$ is a constant and $\$ \notin F$). Given a context $c \in C_F$ and a tree $t \in T_F$, one can build a tree $c[t] \in T_F$ by replacing the (unique) occurrence of \$ in c by the tree t.

Example 1. Let $F_0 = \{a, b\}$, $F_1 = \{g(\cdot)\}$ and $F_2 = \{f(\cdot, \cdot)\}$. Then $t = f(a, g(b)) \in T_F$ (Figure 1(a)), $c = f(a, \$) \in C_F$ (Figure 1(b)) and $c[t] = f(f(a, g(b)), a)$ (Figure 1(c)).
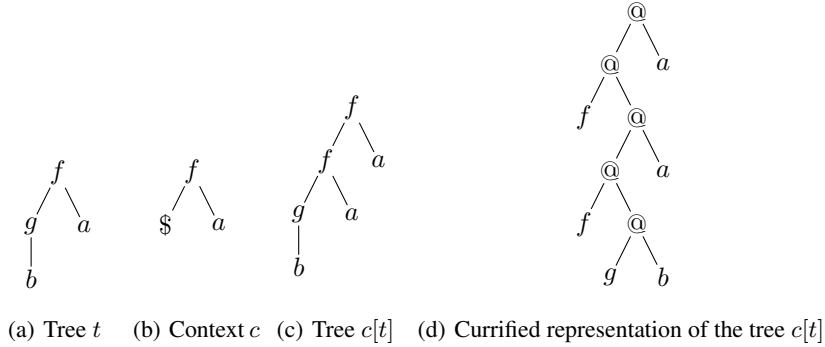
Definition 1. *The* length *of a tree or a context is the number of functional symbols used to define it, including the special symbol* \$.

T_F^k (resp. $T_F^{\geq k}$) will denote the set of trees of length k (resp. length greater or equal than k).

2.2 Tree Series

A *(formal power) tree series* on T_F is a mapping $r : T_F \to \mathbb{R}$. The vector space of all tree series on T_F is denoted by $\mathbb{R}[T_F]$. We denote by $\ell_2(T_F)$ the vector subspace of $\mathbb{R}[T_F]$ of tree series r such that $\sum_{t \in T_F} r(t)^2 < \infty$. This vector subspace is equipped with a dot product $(r, s) = \sum_{t \in T_F} r(t)s(t)$.

Given $r \in \mathbb{R}[T_F]$, a residual of r is a series $s \in \mathbb{R}[T_F]$ such that $s(t) = r(c[t])$ for some $c \in C_F$. This series is denoted $\dot{c}r : t \mapsto r(c[t])$. One defines the set of residuals of

(a) Tree t (b) Context c (c) Tree $c[t]$ (d) Currified representation of the tree $c[t]$

Fig. 1. An example of tree $t = f(a, g(b))$, context $c = f(\$, a)$ and their composition $c[t] = f(f(a, g(b)), a)$, as defined in Example 1. On the right a representation of t over an alphabet with only one binary symbol @ and with the elements of F seen as constant symbols.

r by $\{\dot{c}r | c \in C_F\}$. Let c_i be an enumeration of C_F, and t_j an enumeration of T_F. Given $r \in \mathbb{R}[T_F]$, one defines the (infinite) observation matrix X of r by: $(X)_{i,j} = r(c_i[t_j])$.

$$\begin{pmatrix} r(c_1[t_1]) & \ldots & r(c_1[t_j]) & \ldots \\ \vdots & & \vdots & \\ r(c_i[t_1]) & \ldots & r(c_i[t_j]) & \ldots \\ \vdots & & \vdots & \end{pmatrix}$$

Rational series (series computed by weighted automata) are the series with a finite rank observation matrix. The rank of the observation matrix is the rank of the rational series (i.e. the state number of a minimal automaton computing the series). From this observation, one can define a canonical linear representation of a rational tree series as introduced in [3]. We give here a simpler definition:

Definition 2. *The linear representation of a rational tree series over T_F is given by:*

- *the rank d of the series, and $\{q_1, \ldots q_d\}$ a basis of $\mathbb{R}^d$.*
- *$\tau \in \mathbb{R}^d$.*
- *for each $a \in F_0$, a vector $\underline{a} \in \mathbb{R}^d$.*
- *for $f \in F_2$, a bilinear mapping $\underline{f} \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^d; \mathbb{R}^d)$.*

The linear representation is denoted by $(F, d, \_, \tau)$.

The mapping $\_$ can be inductively extended to a mapping $\_ : T_F \to \mathbb{R}^d$ that satisfies $\underline{f(t_1, t_2)} = \underline{f}(\underline{t_1}, \underline{t_2})$ for any $t_1, t_2 \in T_F$.

Finally, the value of $r(t)$ is given by: $r(t) = \underline{t}^\top \tau$ where $^\top$ denotes the transpose operator.

Example 2. Let $F = \{a, f(\cdot, \cdot)\}$ a ranked alphabet, consider the linear representation $(F, 2, \_, \tau)$ of the series r such that $\{e_1, e_2\}$ is a basis of $\mathbb{R}^2$, $\tau = (1, 0)$ and defined by the following expressions:

$$\underline{a} = \frac{2e_1}{3} + \frac{e_2}{3}, \quad \underline{f}(e_1, e_2) = \frac{e_1}{3} + \frac{2e_2}{3}, \quad \underline{f}(e_i, e_j) = 0 \; for \; (i, j) \neq (1, 2).$$

One has:

$$r(f(a, a)) = \underline{f(a, a)}^\top \tau = \underline{f}(\underline{a}, \underline{a})^\top \tau = \underline{f}(\frac{2e_1}{3} + \frac{e_2}{3}, \frac{2e_1}{3} + \frac{e_2}{3})^\top \tau$$

$$= (\frac{2}{3}\frac{2}{3}\underline{f}(e_1, e_1) + \frac{2}{3}\frac{1}{3}\underline{f}(e_1, e_2) + \frac{1}{3}\frac{2}{3}\underline{f}(e_2, e_1) + \frac{1}{3}\frac{1}{3}\underline{f}(e_2, e_2))^\top \tau$$

$$= (\frac{2}{3}\frac{2}{3}(0, 0) + \frac{2}{3}\frac{1}{3}(\frac{1}{3}, \frac{2}{3}) + \frac{1}{3}\frac{2}{3}(0, 0) + \frac{1}{3}\frac{1}{3}(0, 0)) \cdot \tau = \frac{2}{3^3}.$$

Rational tree series can be equivalently represented by weighted tree automata where the number of states of the automata corresponds to the dimension of the linear representations. Indeed, a tree automaton is a tuple (Q, F, τ, δ) where Q, τ and δ are respectively the set of states, the terminal vector and the transition function. Let $(F, 2, \_, \tau)$ be a linear representation and let $(e_1, \ldots, e_d)$ be a basis of $\mathbb{R}^d$. Let $Q = \{e_1, \ldots, e_d\}$. Any linear relation of the form $\underline{f}(e_i, e_j) = \sum_k \alpha_{i,j}^k e_k$ yields to d transition rules of the form $f(e_i, e_j) \xrightarrow{\alpha_{i,j}^k} e_k$ and $\tau(e_i)$ is set to $\tau^\top e_i$. See [4, 13] for more details.

Example 3. A weighted automaton computing the series in example 2 would be $Q = \{q_1, q_2\}$, $F = \{a, f\}$, δ defined by: $a \xrightarrow{2/3} q_1, a \xrightarrow{1/3} q_2, f(q_1, q_2) \xrightarrow{1/3} q_1, f(q_1, q_2) \xrightarrow{2/3} q_2$ and $\tau(q_1) = 1$, $\tau(q_2) = 0$.

Let $\mathbb{R}[C_F]$ be the set of mappings $s : C_F \to \mathbb{R}$. For $r \in \mathbb{R}[T_F]$ and $t \in T_F$, one can define $\bar{t}r \in \mathbb{R}[C_F]$ by:

$$\bar{t}r(c) = \dot{c}r(t) = r(c[t]).$$

The $\dot{c}r$ correspond to the rows of the observation matrix X and the $\bar{t}r$ to its columns, and one has the following equivalent properties:

1. $r \in \mathbb{R}[T_F]$ has an observation matrix X with finite rank d.
2. The vector subspace of $\mathbb{R}[T_F]$ spanned by $\{\dot{c}r | c \in C_F\}$ has dimension d.
3. The vector subspace of $\mathbb{R}[C_F]$ spanned by $\{\bar{t}r | t \in T_F\}$ has dimension d.

Let us denote by C_n the set of contexts of length lower than n, and let $\sim_{C_n}$ be the equivalence relation over $\mathbb{R}[C_F]$ defined by $f \sim_{C_n} g$ iff $\forall c \in C_n, f(c) = g(c)$. One defines $\mathbb{R}[C_n]$ as the quotient vector space $\mathbb{R}[C_F] / \sim_{C_n}$, equipped with the regular dot product $(f, g) = \sum_{c \in C_n} f(c)g(c)$.

2.3 Rational Distribution and Strong Consistency

Definition 3. *A* rational distribution *(or* rational stochastic tree language*, RSTL) over T_F is a rational series computing a probability distribution.*

In other words, a RSTL is a probability distribution that can be computed by a weighted automaton (or that admits a linear representation). It can be shown that there exists some rational distributions that cannot be computed by any *probabilistic tree automaton*. It is undecidable to know whether a rational series given by a linear representation defines a probability distribution (see [2] for an illustration in the string case).

Definition 4. *A* strongly consistent stochastic tree languages *(or* strongly consistent distribution*) over T_F is a probability distribution over T_F having a bounded average tree size i.e. $\sum_{t \in T_F} p(t)|t| < \infty$.*

It can be shown (see [4]) that, if p is a rational distribution having a bounded average tree size, there exists some constants $0 < C$ and $0 < \rho < 1$ such that:

$$\sum_{t \in T_F^{\geq k}} p(t) \leq C\rho^k.$$

3   Principle of the Algorithm

Let p be rational distribution on T_F (strongly consistent or not). We give first a general algorithm that takes a sample i.i.d. according to p as input. For this purpose, let C_n be the set of contexts of length lower that n and let us make the assumption that $\{\dot{c}p|c \in C_n\}$ and $\{\dot{c}p|c \in C_F\}$ span the same vector subspace of $\mathbb{R}[T_F]$. In other words, we suppose that considering the set C_n is sufficient to get the whole space of residuals.

Let V be the finite dimensional subspace of $\ell^2(T_F)$ spanned by the set $\{\dot{c}p|c \in C_n\}$. V^* will denote the set $\{\bar{t}p|_{C_n}, t \in T_F\} \subset \mathbb{R}[C_n]$ - for convenience we still denote by $\bar{t}p$ the mapping $\bar{t}p|_{C_n}$. Π_V denotes the orthogonal projection over V relatively to the dot product inherited from $\ell_2(T_F)$, and Π_{V^*} denotes the orthogonal projection over V^* relatively to the dot product inherited from $\mathbb{R}[C_n]$. Let S be a sample of N trees independently and identically drawn according to p and let p_S be the empirical distribution on T_F defined from S. V_S denotes the vector subspace of $\ell^2(T_F)$ spanned by $\{\dot{c}p_S|c \in C_n\}$, and V_S^* the subspace of $\mathbb{R}[C_n]$ spanned by $\{\bar{t}p_S|t \in T_F\}$.

We first build from S an estimate $V_{S,d}^*$ of V^* and then we show that $V_{S,d}^*$ can be used to build a linear representation such that its associated rational series approximate the target p. In this section, we implicitly suppose that the dimension d of V^* is known. We will show in the next section how it can be estimated from the data.

3.1   Estimating the Target Space

Let $d > 0$ be an integer. The first step consists in finding the d-dimensional vector subspace $V_{S,d}^*$ of V_S^* that minimizes the distance to $\{\bar{t}p_S|t \in T_F\}$:

$$V_{S,d}^* = \underset{dim(W^*)=d, W^* \subseteq V_S^*}{\arg\min} \sum_{t \in T_F} \|\bar{t}p_S - \Pi_{W^*}(\bar{t}p_S)\|^2.$$

$V_{S,d}^*$ can be computed using principal component analysis.

Let $\{t_j\}$ be an enumeration of T_F, and $\{c_i\}$ be an enumeration of C_n. Let X_S the empirical mean matrix defined by: $(X_S)_{i,j} = p_S(c_i[t_j])$, and let X be the expectation matrix defined by: $(X)_{i,j} = p(c_i[t_j])$.

$V_{S,d}^*$ corresponds to the vector subspace spanned by the d first (normalized) eigenvectors (corresponding to d largest eigenvalues) of the matrix $M_S = X_S X_S^\top$. N_S will denote the matrix $X_S^\top X_S$ which corresponds to the dual problem of the PCA. We will denote by $W^* = \{w_1^*, \ldots, w_d^*\}$ the set of eigenvectors (ordered by decreasing eigenvalues) of M_S, and by $W = \{w_1, \ldots, w_d\}$ the corresponding eigenvectors of N_S. W^* is the matrix with the vectors $\{w_1^*, \ldots, w_d^*\}$ as columns - this matrix corresponds to the projection operator Π_{V^*}, while W is the matrix with vectors $\{w_1, \ldots, w_d\}$ as columns, because both W and W^* are orthonormal.

Let $\lambda_1, \ldots, \lambda_d$ be the associated singular values; they also are the square roots of the eigenvalues of M_S.

We recall here the relationships between the w_i and w_i^* eigenvectors: $X_S w_i = \lambda_i w_i^*$ and $X_S^\top w_i^* = \lambda_i w_i$. In particular,

$$M_S w_i^* = X_S X_S^\top w_i^* = \lambda_i X_S w_i = \lambda_i^2 w_i^*.$$

3.2   Building the Linear Representation From the Dual Space

The eigenvectors found in the previous section form the basis of the residual space. In order to complete the linear representation, we now need to define, in the basis $\{w_1^*, \ldots, w_d^*\}$, the terminal vector τ, the mapping $\_$ for the constant symbols $\underline{a}$ and the bi-linear operator $\underline{f}$.

The idea is to identify, for any tree t, the mapping $\bar{t}p_S$ to its projection on the space spanned by W^*, that is $W^* W^{*\top} \bar{t} p_S$. We shall see in next section that this identification leads to a bounded error, decreasing as the size of the sample grows.

- The vector space is the space spanned by W^*.
- For each $a \in F_0$, $\underline{a} = W^{*\top} \bar{a} p_S$.

In order to define $\underline{f}$, we use a known relation between eigenvectors of the standard and dual PCA: $w_i^* = \sum_k \frac{(w_i)_k}{\lambda_i} \bar{t}_k p_S$. We use the bilinearity of $\underline{f}$ to obtain:

- $\underline{f}(w_i^*, w_j^*) = \sum_{1 \le k,l \le d} \frac{(w_i)_k (w_j)_l}{\lambda_i \lambda_j} W^{*\top} \overline{f(t_k, t_l)} p_S$.
- $\tau_i = w_i^*(\$)$, corresponding to the terminal weight of a tree in a bottom-up process.
- Finally, $r(t) = \underline{t}^\top . \tau$.

The different steps of the algorithm are described in Algorithm 1.

4   Consistency

Let us consider the observation matrix X defined by: $X_{ij} = p(c_i[t_j])$, where c_i and t_j are respectively contexts and trees. Let S be a sample of size N i.i.d. from p. X_S is defined as the empirical observation matrix built from the empirical distribution p_S. In

Data: A sample S of trees in T_F i.i.d. according to a distribution p, a dimension d and a set of contexts C_n.
Result: A linear representation A of a tree series $(F, d, \_, \tau)$.
Let X the matrix defined by $X[i,j] = p_S(c_i[t_j])$;
$M = XX^\top$ /* variance-covariance matrix */;
$(\lambda_i, w_i^*, w_i) \leftarrow$ square roots of eigenvalues of M in decreasing order and corresponding eigenvectors, and eigenvectors in the dual;
Let $w_1^*, \ldots, w_d^*$ be the eigenvectors corresponding to the d largest eigenvalues and let $W^* = [w_1^*, \ldots, w_d^*]$ be the matrix having the vectors w_i^* as columns
Let $\_$ be the operator defined by:
foreach $f \in \mathcal{F}$ **do**
> **if** $a \in \mathcal{F}_0$ **then** $\underline{a} = {W^*}^\top \bar{a} p_S$;
> **if** $f \in \mathcal{F}_2$ **then** $\underline{f}(w_i^*, w_j^*) = \sum_{1 \le k, l \le d} \frac{(w_i)_k (w_j)_l}{\lambda_i \lambda_j} {W^*}^\top \overline{f(t_k, t_l)} p_S$;

end
$(\tau)_i = w_i^*(\$)$;
return $A = (F, d, \_, \tau)$;

Algorithm 1: Building a linear representation corresponding to a sample S and a dimension d.

this section, we will bound the difference between those two matrices, and show how it induces a bound for the convergence of the singular values and on the distance between the estimate and the target distribution for a tree t.

First, here is a simple result straightforward from the properties of empirical mean:

Lemma 1. *Let p a probability distribution over T_F, and p_S its empirical estimate from a sample of size N drawn i.i.d. from p, one has*

$$\mathbf{E}(\|p_S - p\|_2^2) = \sum_{t \in T_F} \mathbf{E}((p_S(t) - p(t))^2) = \sum_{t \in T_F} \frac{p(t)(p(t)-1)}{N} \le \frac{1}{N}.$$

Let C_n be the set of contexts of length lower or equal than n, it can easily be shown that:

Lemma 2. *Let t be a tree. There is at most n contexts in C_n such that $t = c[t']$ for some tree t'.*

The previous lemma helps us to bound the occurrence number of a tree in the matrix X, which will allow us to use some concentration inequality to bound the error over X. One denotes $\|\|_F$ as the Frobenius norm on matrices, and $\Delta_X = \|X - X_S\|_F$.

Lemma 3. *Let X be a probability observation matrix restricted to the contexts belonging to C_n. Let X_S the empirical estimator of X from a sample S of size N. Then, one has with probability at least $1 - \delta$ ($\delta > 0$):*

$$\Delta_X = \|X - X_S\|_F \le \sqrt{\frac{n}{N}} \left(1 + \sqrt{\log(\frac{1}{\delta})}\right).$$

Proof. This proof uses a construction similar to the proof of Proposition 19 in [8]. Let z be a discrete random variable that takes values in T_F. Let X be a probability observation matrix built from a set C_n of contexts as lines and trees from T_F as columns. One estimates X from N i.i.d. copies of z_i of z ($i = 1, \ldots, N$).

One associates to each variable z_i a matrix X_i indexed by contexts of C_n and trees of T_F such that

$$X_i[j, k] = 1 \text{ if } z_i = c_j[t_k] \text{ and } 0 \text{ otherwise.}$$

From Lemma 2, X_i has at most n non null entries.

The empirical estimate of X is $X_S = \frac{1}{N} \sum_{i=1}^{N} X_i$. Our objective is to bound $\|X_S - X\|_F$.

Let S' be a sample that differs from S on at most one example z'_k.

Then,

$$\left| \|X_S - X\|_F - \|X_{S'} - X\|_F \right| \leq \|X_S - X_{S'}\|_F \leq \sqrt{\frac{2n}{N}}.$$

From McDiarmid inequality [14], one obtains:

$$Pr(\|X_S - X\|_F \geq \mathbf{E}(\|X_S - X\|_F) + \epsilon) \leq e^{-\frac{N}{n}\epsilon^2}.$$

By Lemma 1 and Lemma 2 and by using Jensen's inequality, it can be proved that $\mathbf{E}(\|X_S - X\|_F) \leq \sqrt{\frac{n}{N}}$. By fixing $\delta = e^{-\frac{N}{n}\epsilon^2}$, one gets the result. □

4.1 Singular Values Convergence

We use the previous result to show how one can assess the correct dimension of the target space. We will first recall some known result. Given an observation matrix X of rank d in the target space, and given its empirical estimate X_S, we can rewrite X_S as a sum $X + E$ where E models the sampling error. We have the following result from [15].

Lemma 4. *(Theorem 4.11 in [15]). Let $X \in \mathbb{R}^{m \times n}$ with $m \geq n$, and let $X_S = X + E$. If the singular values of X and X_S are $(\lambda_1 > \ldots > \lambda_n)$ and $(\lambda_{S,1} > \ldots > \lambda_{S,n})$ respectively, then*

$$|\lambda_{S,i} - \lambda_i| \leq \|E\|_2, i = 1, \ldots, n.$$

Applied to our situation, this provides a valid way to assess the target dimension: let d be the rank of the target rational series, X_S be the observation matrix deduced from a sample S, $|S| = N$.

Theorem 1. *Let Λ be the set of singular values of X_S. Let Λ_s be the subset of singular values of X_S greater than s. For a given confidence parameter δ, let $d' = |\Lambda_s|$ for $s = \sqrt{\frac{n}{N}}(1 + \sqrt{\log(\frac{1}{\delta})})$. With probability greater than $1 - \delta$, one has $d \geq d'$.*

Proof. Straightforward from Lemma 3 and Lemma 4: with probability greater than $1 - \delta$, the singular values in Λ_s match non-zeros singular values from the target observation matrix X. □

Theorem 2. *Let λ_d the smallest non-zero eigenvalue of X. Let Λ be the set of singular values of X_S. Let Λ_s be the subset of singular values of X_S greater than s. For a given confidence parameter δ, let $d' = |\Lambda_s|$ for $s = \sqrt{\frac{n}{N}}(1 + \sqrt{\log(\frac{1}{\delta})})$. Suppose that*

$$N > \frac{4n}{\lambda_d^2}\left(1 + \sqrt{\log(\frac{1}{\delta})}\right)^2$$

Then, with probability greater than $1 - \delta$, one has $d = d'$.

Proof. The condition $N > \frac{4n}{\lambda_d^2}(1 + \sqrt{\log(\frac{1}{\delta})})^2$ implies that $s < \frac{\lambda_d}{2}$, thus the corresponding singular value $\lambda_{S,d}$ from X_S satisfies $\lambda_{S,d} > 2s - \|X - X_S\|_2$. This quantity is greater than s with probability at least $1 - \delta$. $\qquad\square$

4.2 Bounds for the Estimation Error

We suppose here that the correct dimension has been found. We will not provide exact bounds, but only asymptotic bounds, and we will often use the equivalence between norms of vectors and matrices - since the vector spaces considered are finite-dimensional. Let us first introduce some notations corresponding to errors over the objects handled by our algorithm:

- $\Delta_x = \max \|x - x_S\|_2$ with x (resp. x_S) a row or a column of the observation matrix X (resp. X_S).
- $\Delta_v = \max \|w - w_S\|_2$ with w (resp. w_S) a left singular vector of the observation matrix X (resp. X_S).
- $\Delta_\lambda = \max \|\lambda - \lambda_S\|_2$ with λ (resp. λ_S) a singular value of the observation matrix X (resp. X_S).
- $\Delta_\Pi = \|WW^T - W_S W_S^T\|_F$ with W (resp. W_S) the d first singular vectors of the observation matrix X (resp. X_S).

Lemma 5. $\Delta_\lambda < \Delta_X$ *and* $\Delta_x < \Delta_X$.

Proof. Straightforward from Lemma 4 and the norm relation $\|\|_2 \leq \|\|_F$ for the first inequality, and the definition of Δ_x and Δ_X for the second. $\qquad\square$

The following corollary gives an asymptotic bound on the error of the covariance matrix used to compute the eigenvectors.

Corollary 1. $\Delta_M = \|M - M_S\|_F$. *One has* $\|M - M_S\|_F \leq O(\Delta_X)$

Proof. One has $\|M - M_S\|_F \leq \|XX^\top - XX_S^\top + XX_S^\top - X_S X_S^\top\|_F \leq (\|X\|_F + \|X_S^\top\|_F)\Delta_X$. Thus:

$$\Delta_M \leq \Delta_X(2\|X\|_F + \Delta_X).$$

$\qquad\square$

In order to provide asymptotic bounds on the other errors, we need to introduce some known results about eigenvectors and PCA from [16]. Let A be a symmetric positive Hilbert-Schmidt operator with positive eigenvalues[1] $\lambda_1^2 > \cdots > \lambda_d^2 > 0$. $\delta_r = \frac{1}{2}(\lambda_r^2 - \lambda_{r+1}^2)$, and let $\tilde{\delta}_r = \inf(\delta_r, \delta_{r-1})$. Let B be a symmetric positive Hilbert-Schmidt operator such that $\|B\|_F < \tilde{\delta}_r/2$ and that $\|B\|_F < \delta_d/2$. The results from [16] provide error bounds on projection operators and eigenvectors. In our framework, A corresponds to the covariance matrix M and $A + B$ to the empirical one M_S. Let W (resp. W_S) be the matrix of the d first eigenvectors of A (resp. $A + B$), and w_r (resp. $w_{S,r}$) the corresponding r-th eigenvector, we have the following results.

Lemma 6. *(Theorem 2 - remark of [16])* $\|w_r - w_{S,r}\|_2 \leq \frac{2\|B\|_F}{\tilde{\delta}_r}$.

Theorem 3. *(Theorem 3 of [16])* $\|WW^\top - W_S W_S^\top\|_F \leq \frac{\|B\|_F}{\delta_d}$.

We are now able to provide asymptotic bounds for the two remaining errors.

Lemma 7. *One has $\Delta_v = O(\Delta_X)$ and $\Delta_\Pi = O(\Delta_X)$.*

Proof. By using respectively Lemma 6 and Theorem 3, and from Corollary 1, one has

$$\Delta_v \leq \frac{4\Delta_X(2\|X\|_F + \Delta_X)}{\tilde{\delta}_d} = O(\Delta_X)$$

and

$$\Delta_\Pi \leq \frac{\Delta_X(2\|X\|_F + \Delta_X)}{\tilde{\delta}_d} = O(\Delta_X).$$

$\square$

Let us denote $\lambda = \inf_{1\ldots d} \lambda_i$. We will now study some errors on the parameters of the linear representation built by our algorithm. Let $p = (F, d, \_, \tau)$ be the target linear representation equipped with the basis $\{w_1^*, \ldots, w_d^*\}$ and let $r_S = (F, d, \_S, \tau_S)$ the linear representation equipped with the basis $\{w_{S,1}^*, \ldots, w_{S,d}^*\}$ found by our algorithm from a sample S. Let us define the following error bounds on the coefficients:

- $\Delta_\tau = \sup_i(\tau - \tau_S)_i$,
- $\Delta_{\underline{a}} = \sup_i(\underline{a} - \underline{a}_S)_i$ for $a \in F_0$,
- $\Delta_{\underline{f}} = \sup_{i,j,k}(\underline{f}(w_j^*, w_k^*) - \underline{f}_S(w_{S,j}^*, w_{S,k}^*))_i$.

One can check the following lemma.

Lemma 8.

$$\Delta_\tau \leq dO(\Delta_v) \leq 2d\|X\|_F O(\Delta_X),$$
$$\Delta_{\underline{a}} \leq 2O(\Delta_X),$$
$$\Delta_{\underline{f}} \leq \frac{O(\Delta_X)}{\lambda}\left[\|w_i w_j^\top\|_F(2 + 2\frac{\|X\|_F}{\delta} + 8\frac{\|X\|_F}{\delta}(\|w_i\|_1 + \|w_j\|_1))\right].$$

---

[1] Recall that according to our notation the λ_i denote singular values.

All the mappings considered through the algorithm are continuous, thus the mapping deduced from the algorithm converges pointwisely towards the target distribution as Δ_X tends to zero. We provide then the following result which gives a bound for the estimation error.

Theorem 4. *Let p be a rational distribution with rank d. Let n be the maximum length of a context used in the algorithm. If S is a sample i.i.d. from p, let r_S be the mapping deduced from the algorithm. There exists C such that for any $0 < \delta < 1$, for any tree t of length k, if S is an i.i.d. sample of size N then, with confidence at least $1 - \delta$, one has:*

$$|r_s(t) - p(t)| < Ckd^{2k}\sqrt{\frac{n}{N}\log(\frac{1}{\delta})}.$$

Proof. Let us prove the statement by induction on k. Let us denote Δ_k a bound for the error made for the estimate the coefficient of $\underline{t}$ for a tree t of height k. One has:

$$\Delta_1 = \Delta_{\underline{a}} = O(\Delta_X)$$

Using $\underline{f}(u,v) = \sum_{1 \leq i \leq d}\sum_{1 \leq i \leq d}(\underline{u})_i(\underline{v})_j\underline{f}(w_i^*, w_j^*)$, with $|u| + |v| = k - 1$, one has:

$$\Delta_k = O(d^2(|u|d^{2|u|}\Delta_X + |v|d^{2|v|}\Delta_X + \Delta_X)) \leq O(kd^{2k}\Delta_X) = O(kd^{2k}\sqrt{\frac{n}{N}\log(\frac{1}{\delta})}).$$

Then, since $p(t) = \underline{t}^T.\tau$ and $r_s(t) = \underline{t}_S^T.\tau_S$, one has the conclusion. □

4.3 Strongly Convergent Case

In the case of strongly consistent distribution, one does not need to consider a finite set of contexts to perform the algorithm: one has, with confidence greater than $1 - \delta$, $|t| < \frac{\log(2C/\delta)}{\log(1/\rho)}$. One can provide a bound result for this special case.

Theorem 5. *Let p be a* strongly consistent *rational distribution with rank d. Let S be a sample i.i.d. from p, let r_S be the mapping deduced from the algorithm. There exists C such that for any $0 < \delta < 1$, for any tree t of length k, if S is an i.i.d. sample of size N then, with confidence at least $1 - \delta$, one has:*

$$|r_s(t) - p(t)| < Ckd^{2k}\sqrt{\frac{\log(N)}{N}log^2(\frac{1}{\delta})}.$$

Proof. One bounds the length of a tree drawn by p: with a confidence greater than $1 - \delta/2N$,

$$|t| < \frac{\log(2NC/\delta)}{\log(1/\rho)}.$$

Thus, with confidence $1 - \delta/2$, S has only trees of length lower than $\frac{\log(2NC/\delta)}{\log(1/\rho)}$. By replacing n in the previous result, one obtains the conclusion. □

5   Illustration

In order to illustrate the algorithm, we consider the distribution p defined by tree series of Example 2.

To study the behavior our algorithm, we consider an observation matrix 5×5, built on the set of trees $T = \{t_1, t_2, t_3, t_4, t_5\}$, where $t_1 = a$, $t_2 = f(a, a)$, $t_3 = f(a, f(a, a))$, $t_4 = f(f(a, a), a)$, $t_5 = f(f(a, a), f(a, a))$ and the set of contexts

$$C = \{\$, f(a, \$), f(\$, a), f(f(a, a), \$), f(\$, f(a, a))\}.$$

We generate i.i.d. samples from p of different sizes containing respectively 10^3, 10^4, 10^5 and 10^6 trees. On Figure 2, we show the different eigenvalues (square of singular values) in decreasing order obtained from the different samples.

We can observe that the convergence of the computed series towards the target value, and the convergence of singular values, is closely $O(\frac{1}{\sqrt{|S|}})$ (in average values).
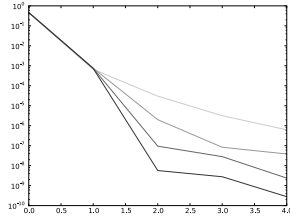
We also compared the average standard deviation of the probabilities of the trees in T obtained with our model, with rank 2 learned from the different learning samples, with the theoretical standard deviation of the classical probability (binomial) estimator. We have that $p(t_1) \simeq 0.6666$, $p(t_2) \simeq 0.0741$, $p(t_3) \simeq 0.0329$, $p(t_4) \simeq 0.0082$ and $p(t_5) \simeq 0.0037$. The values obtained are shown on Figure 3. The estimated standard deviation is, in majority (21/25), lower than the theoretical standard deviation of a the binomial estimator: The algorithm seems to work better than the simple frequency estimator for the task of density estimation.

Now, we consider the problem of the dimension estimate. The second singular value $\lambda_2 \sim 7.74 \cdot 10^{-3}$. Using the bound Δ_X to estimate the correct dimension (Theorem 2), we can estimate that:

– For $N = 10^6$, the rank 2 is found with a parameter $\delta \sim 0.59$ (confidence 0.41).
– For $N = 2 \cdot 10^6$, the rank 2 is found with a parameter $\delta \sim 0.12$ (confidence 0.88).
– For $N = 3 \cdot 10^6$, the rank 2 is found with a parameter $\delta \sim 0.02$ (confidence 0.98).

6   Conclusion and Discussion

We have studied the problem of learning an unknown distribution p from finite independently and identically drawn samples. We have proposed a new approach for identifying rational distributions on trees, or rational stochastic tree languages. Most classical inference algorithms in probabilistic grammatical inference build an automaton or a grammar iteratively from a sample S. Starting from an automaton composed of only one state, then they have to decide whether a new state must be added to the structure. This iterative decision relies on a statistical test with a known drawback: as the structure grows, the test relies on fewer and fewer examples. Instead of this iterative approach, we tackle the problem globally and our algorithm computes in one step the space needed to build the output automaton. That is, we have reduced the problem set

t	σ_{10^2}	σ_{10^3}	σ_{10^4}	σ_{10^5}	σ_{10^6}
t_1	$3.76.10^{-2}$	$1.23.10^{-2}$	$3.50.10^{-3}$	$1.16.10^{-3}$	$4.12.10^{-4}$
	$4.71.10^{-2}$	*$1.49.10^{-2}$*	*$4.71.10^{-3}$*	*$1.49.10^{-3}$*	*$4.71.10^{-4}$*
t_2	$2.04.10^{-2}$	$6.77.10^{-3}$	$1.94.10^{-3}$	$7.30.10^{-4}$	$2.36.10^{-4}$
	$2.62.10^{-2}$	*$8.28.10^{-3}$*	*$2.62.10^{-3}$*	*$8.28.10^{-4}$*	*$2.62.10^{-4}$*
t_3	$1.63.10^{-2}$	$6.70.10^{-3}$	$2.13.10^{-3}$	$6.95.10^{-4}$	$1.95.10^{-4}$
	$1.78.10^{-2}$	*$5.64.10^{-3}$*	*$1.78.10^{-3}$*	*$5.64.10^{-4}$*	*$1.78.10^{-4}$*
t_4	$9.01.10^{-3}$	$2.23.10^{-3}$	$6.93.10^{-4}$	$2.20.10^{-4}$	$5.73.10^{-5}$
	$9.03.10^{-3}$	*$2.86.10^{-3}$*	*$9.03.10^{-4}$*	*$2.86.10^{-4}$*	*$9.03.10^{-5}$*
t_5	$4.90.10^{-3}$	$1.51.10^{-3}$	$4.52.10^{-4}$	$1.46.10^{-4}$	$3.90.10^{-5}$
	$6.04.10^{-3}$	*$1.91.10^{-3}$*	*$6.04.10^{-4}$*	*$1.91.10^{-4}$*	*$6.04.10^{-5}$*

Fig. 2. Curves of eigenvalues - square of singular values - (in logarithmic scale) for sample size of 10^3, 10^4, 10^5 and 10^6 trees (the lightest to the darkest).

Fig. 3. Average standard deviation of trees in T measured from the 2-dimensional model learned on samples of size 10^2, 10^3, 10^4, 10^5 and 10^6. The standard deviation of the theoretical binomial estimator is indicated in *italics.*

in the classical probabilistic grammatical inference framework to a classical optimization problem. This point offers the interesting opportunity to apply classical results in statistical machine learning theory to probabilistic grammatical inference.

We have provided three types of results. First, we have given a result for convergence of eigenvalues which can be used for the estimation of the dimension of the target vector space, which is a crucial point in probabilistic grammatical inference and may allow to avoid costly cross-validation procedures. Second, we have provided error bounds for the convergence of the parameters of a linear representation. We have finally obtained pointwise convergence results for the probability estimate of a tree.

One perspective would then to obtain an ℓ_1-convergence, probably restricted to the case of strongly consistent stochastic tree languages, and to obtain tighter bounds. We finally need to experimentally study and compare our approach to existing ones on real data, this is a work in progress. Another perspective would consist in introducing non linearity via the kernel PCA technique developed in [17] and by the Hilbert space embedding of distributions proposed in [18, 19].

Acknowledgement

References

1. Carrasco, R., Oncina, J., Calera-Rubio, J.: Stochastic inference of regular tree languages. Machine Learning **44**(1/2) (2001) 185–197
2. Denis, F., Esposito, Y.: On rational stochastic languages. Fundamenta Informaticae **86** (2008) 41–77

3. Denis, F., Habrard, A.: Learning rational stochastic tree languages. In: Proc. of 18th conf. on Algorithmic Learning Theory. Volume 4754 of LNCS., Springer-Verlag (2007) 242–256

4. Denis, F., Gilbert, E., Habrard, A., Ouardi, F., Tommasi, M.: Relevant representations for the inference of rational stochastic tree languages. In: Grammatical Inference: Algorithms and Applications, 9th International Colloquium, Springer (2008) 57–70

5. Bailly, R., Denis, F., Ralaivola, L.: Grammatical inference as a principal component analysis problem. In: Proceedings of the 26th International Conference on Machine Learning, Montréal, Canada, Omnipress (June 2009) 33–40

6. Clark, A., Costa Florêncio, C., Watkins, C.: Languages as hyperplanes: grammatical inference with string kernels. In: Proceedings of the European Conference on Machine Learning (ECML). (2006) 90–101

7. Hsu, D., Kakade, S., Zhang, T.: A spectral algorithm for learning hidden markov models. In: Proceedings of COLT'2009, Springer (2009)

8. Hsu, D., Kakade, S., Zhang, T.: A spectral algorithm for learning hidden markov models. Technical report, Arxiv archive (2009) http://arxiv.org/abs/0811.4413.

9. Denis, F., Esposito, Y.: Learning classes of probabilistic automata. In: Proc. 17h Conference on Learning theory (COLT'04). Volume 3120 of LNCS., Springer (2004) 124–139

10. Comon, H., Dauchet, M., Gilleron, R., Jacquemard, F., Lugiez, D., Löding, C., Tison, S., Tommasi, M.: Tree automata techniques and applications. Available on: http://tata.gforge.inria.fr/ (2007) release October, 12th 2007.

11. Berstel, J., Reutenauer, C.: Recognizable formal power series on trees. Theorical computer science **18** (1982) 115–148

12. Högberg, J., Maletti, A., Vogler, H.: Bisimulation minimisation of weighted automata on unranked trees. Fundam. Inform. **92**(1-2) (2009) 103–130

13. Borchardt, B.: The Theory of Recognizable Tree Series. PhD thesis, TU Dresden (2004)

14. McDiarmid, C.: On the method of bounded differences. In: Surveys in Combinatorics. Cambridge University Press (1989) 148–188

15. Stewart, G., Sun, J.G.: Matrix Perturbation Theory. Academic Press (1990)

16. Zwald, L., Blanchard, G.: On the convergence of eigenspaces in kernel principal component analysis. In: Proceedings of NIPS'05. (2006)

17. Shawe-Taylor, J., Cristianini, N., Kandola, J.: On the concentration of spectral properties. In: Proc. of NIPS. Volume 14., MIT Press (2001) 511–517

18. Smola, A., Gretton, A., Song, L., Schölkopf, B.: A hilbert space embedding for distributions. In: 18th International Conference on Algorithmic Learning Theory, Springer (2007) 13–31

19. Song, L., Boots, B., Saddiqi, S., Gordon, G., Smola, A.: Hilbert space embeddings of Hidden Markov Models. In: Proceedings of ICML 2010